

# Untangling influences of hydrophobicity on protein sequences and structures

Mehdi Yahyanejad,<sup>1,2</sup> Christopher B. Burge,<sup>2</sup> and Mehran Kardar<sup>1</sup>

<sup>1</sup>*Department of Physics, Massachusetts Institute of Technology, Cambridge, MA 02139*

<sup>2</sup>*Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139*

We fit the Fourier transforms of solvent accessibility and hydrophobicity profiles of a representative set of proteins to a joint multi-variable Gaussian. This allows us to separate the intrinsic tendencies of sequence and structure profiles from the interactions that correlate them; for example, the  $\alpha$ -helix periodicity in sequence hydrophobicity is dictated by the solvent accessibility of structures. The distinct intrinsic tendencies of sequence and structure profiles are most pronounced at long periods, where sequence hydrophobicity fluctuates more, while solvent accessibility fluctuations are less than average. Interestingly, correlations between the two profiles can be interpreted as the Boltzmann weight of the solvation energy at room temperature.

How the sequence of amino acids determines the structure and function of the folded protein remains a challenging problem. It is known that hydrophobicity is an important determinant of the folded state; hydrophobic monomers tend to be in the core, and polar monomers on the surface [1, 2, 3, 4]. Several studies have examined the correlations in the hydrophobicity of amino-acids along the protein chain [5, 6, 7, 8], which are in secondary structure prediction [9], and in the design of good folding sequences [10]. Naturally, sequence correlations arise from locations of the amino-acids in the folded protein structure, and are best interpreted in conjunction with solvent accessibility profiles (which indicate how exposed a particular amino-acid is to water in a specific structure). For example, Eisenberg *et al* [3] note that for secondary structures lying on the protein surface, which have a strong periodicity in their solvent accessibility, hydrophobicity profiles also exhibit the period of the corresponding  $\alpha$  helix or  $\beta$  strand. Constraints from forming compact structures induce strong correlations in the solvent accessibility profile [7, 11, 12, 13], which should in turn induce similar correlations in the hydrophobicity profiles. It is desirable to quantify and separate the resulting correlations in protein sequences and structures.

In this paper, we aim for a unified treatment of hydrophobicity and solvent accessibility profiles, and the interactions between them. The *sequence* of each protein is represented by a profile  $\{h_i\}$ , where  $h_i$  is a standard measure of the hydrophobicity of the  $i$ -th amino-acid along the backbone [14]. Its *structure* has a profile  $\{s_i\}$  for  $i = 1, 2, \dots, N$ , where  $s_i$  is a measure of the exposure of the amino-acid to water in the folded structure [2]. While we do not expect perfect correlations between these profiles, we can inquire about the statistical nature of these correlations, and in particular whether they are diminished or enhanced at different periods. To this end, we employ the method of Fourier transforms, and examine the statistics of the resulting amplitudes  $\{\tilde{h}_q, \tilde{s}_q\}$ , and power spectra  $\{|\tilde{h}_q|^2, |\tilde{s}_q|^2\}$ , for a database of 1461 non-homologous proteins. In a sense, this can be regarded as extending the work of Eisenberg *et al* [3] who explore correlations between hydrophobicity and solvent

independent of specific locations along the backbone. Of course, the use of Fourier analysis is by no means new, and has for example been employed to study hydrophobicity profiles [3, 5, 15, 16]. However, we are not aware of its use as a means of correlating sequence and structure profiles.

Our results suggest that the hydrophobicity and solvent accessibility profiles are well approximated by a joint Gaussian probability distribution. This allows us to obtain the *intrinsic* correlations in the hydrophobicity profile, as distinct from correlations induced by solvent accessibility. For example, the  $\alpha$ -helix periodicity in hydrophobicity profiles is shown to be induced by the corresponding periodicity in the solvent accessibility profiles. We also find that at long wavelengths the two profiles have different intrinsic characteristics: solvent accessibility profiles are positively correlated while hydrophobicity profiles are anti-correlated. Interestingly, the coupling between the two profiles is independent of wave-number, and hence can be interpreted as the Boltzmann weight of the solvation energy. The corresponding temperature is close to room temperature, consistent with the “mean” temperature estimated in previous work from the frequencies of occurrence of amino acid residues in the core and on the surface [17, 18].

For our protein data set, we selected 2200 representative chains from the Dali/FSSP database. Any two protein chains in this set have more than 25 percent structural dissimilarity. We removed all the multi-domain chains by using the CATH domain definition database, leaving 1461 protein chains [4, 19, 20]. The hydrophobicity profiles,  $\{h_i\}$ , were generated from the sequence of amino-acids using the experimentally measured scale of Fache and Pliska [21] (in units of kcal/mol). We used the *relative solvent accessibility* reported by NACCESS [22] to generate solvent accessibility profiles  $\{s_i\}$ . (The relative solvent accessibility is the ratio of the solvent accessibility of a residue to the solvent accessibility of that residue in an extended tripeptide ALA-X-ALA for each amino acid type X.) We then computed the cor-

responding Fourier components as

$$\tilde{s}_q = \frac{1}{\sqrt{N}} \sum_{j=1}^N e^{iqj} \left( s_j - \frac{\sum_{j=1}^N s_j}{N} \right), \quad (1)$$

where  $q = 2\pi\alpha/N$ , with  $\alpha = 0, 1, \dots, N-1$ , and similarly for  $\tilde{h}_q$ . (The average values were subtracted to remove the DC component in the Fourier transform.)

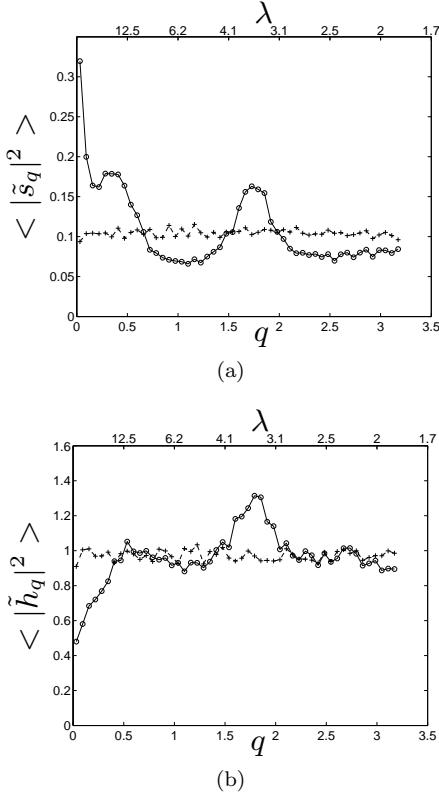


FIG. 1: Power spectra from averaging over 1461 proteins, for (a) solvent accessibility (there are no units for  $|\tilde{s}_q|^2$ , since it based on accessibility relative to solution); (b) hydrophobicity (the units of  $|\tilde{h}_q|^2$  are  $(k \text{ cal/mole})^2$ ). The plus signs in each case are obtained from random permutation of the sequences.

Our results for the power spectra of solvent accessibility and hydrophobicity profiles are indicated respectively in Figs. 1(a) and 1(b) ( $q$  is related to the periodicity  $\lambda$  through  $\lambda = \frac{2\pi}{q}$ ). A prominent feature of both plots is the peak at the  $\alpha$ -helix periodicity  $\lambda = 3.6$  [3]. Its presence in the solvent accessibility spectrum indicates that solvation energy plays a role in the spatial arrangement of  $\alpha$  helices when they lie on the surface, the hydrophobic monomers are more likely to be exposed to the solvent. [3].

We would like to untangle correlations between the two profiles, so as to determine their intrinsic tendencies, by finding a joint probability distribution  $P(\{s_i\}, \{h_i\})$ . Clearly this cannot be decomposed as a product of contributions from different sites  $i$ , as neighboring components such as  $s_i$  and  $s_{i+1}$ , are highly correlated. We antici-

pate that the Fourier components for different  $q$  are independently distributed (i.e.  $P(\{\tilde{h}_q, \tilde{s}_q\}) = \prod_q p(\tilde{h}_q, \tilde{s}_q)$ ) for the following reasons: (i) For the subgroup of cyclic proteins [23] the index  $i$  is arbitrary, and the counting can start from any site. The invariance under relabeling then implies that the probability can only depend on  $i - j$ , and hence separable into independent Fourier components. This exact result does not hold for open proteins because of end effects, but should be approximately valid for long sequences when such effects are small. (ii) Numerical analysis of a lattice model of proteins in Ref. [12] confirms the exact decomposition into Fourier modes for cyclic structures, and its robustness even for open structures of only  $N = 36$  monomers. To test this hypothesis, we examine all possible covariances involving  $\{\tilde{h}_q, \tilde{s}_q\}$  for different  $q$ . Note that the Fourier amplitudes are complex (i.e.  $\tilde{s}_q = \Re s_q + i\Im s_q$ , and similarly for  $\tilde{h}_q$ ), and hence there are  $4 \times 4$  covariance plots, such as in Fig. 2(a)) for the covariance of  $\Re s_q$  with itself. In all cases we find that the diagonal terms are small; the only exceptions are at small  $q$  where we expect end effects to be most pronounced.

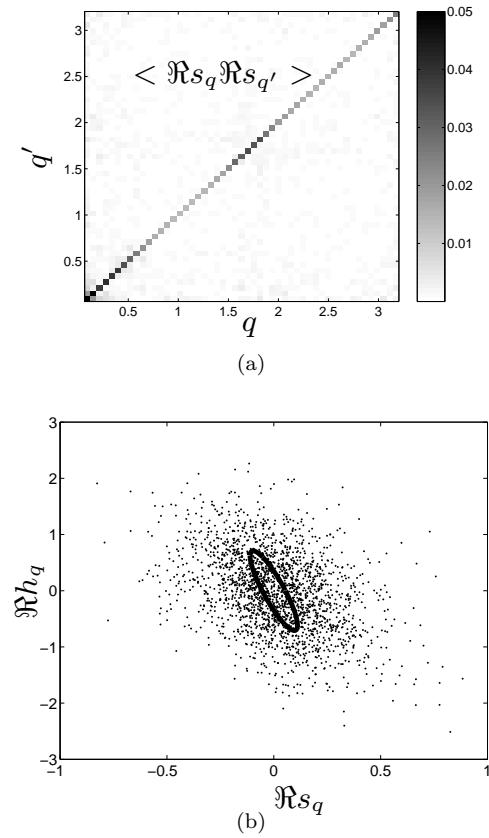


FIG. 2: (a) Covariance of  $\Re s_q$  with  $\Re s_{q'}$ . There is very little correlation between off-diagonal terms. (b) Scatter plot of  $\Re h_q$  versus  $\Re s_q$  for  $q = 0.90$ , and the half-width half-maximum locus of a Gaussian fit (solid line).

One can make a similar case for the independence of the real and imaginary components at a given  $q$ . (For cyclic structures the phase is arbitrary.) The real (imaginary) components are, however, correlated as illustrated by the scatter plot of  $(\Re h_q, \Im s_q)$  for  $q = 0.9$  in Fig. 2(b). We made similar scatter plots for different values of  $q$  in the interval 0 to  $\pi$ , with similar results which were well fitted to Gaussian forms. Based on these results, we describe the joint probability distribution in Fourier space by the multivariate Gaussian form

$$P(\{\tilde{h}_q, \tilde{s}_q\}) = \prod_q \exp \left[ -\frac{(\Re s_q)^2}{2A_q} - \frac{\Re s_q \Re h_q}{B_q} - \frac{(\Re h_q)^2}{2C_q} \right] \\ \times \exp \left[ -\frac{(\Im s_q)^2}{2A'_q} - \frac{\Im s_q \Im h_q}{B'_q} - \frac{(\Im h_q)^2}{2C'_q} \right], \quad (2)$$

with the parameters plotted in Fig. 3. If the probabilities depend only on the separation  $i - j$  between sites, the real and imaginary Fourier amplitudes should follow the same distribution. In our fits we allowed the corresponding parameters to be different to obtain a measure of the accuracy of the model and the fitting procedure. As indicated in Fig. 3 the resulting values are quite close, differing by less than 5%.

We interpret  $\{A_q\}$  and  $\{C_q\}$  as measures of *intrinsic* tendencies of hydrophobicity and surface exposure profiles, while  $\{B_q\}$  indicates the strength of the interactions that correlate them. In the absence of any such interactions,  $\{A_q\}$  and  $\{C_q\}$  would be the same as the power spectra in Fig. 1. With this in mind, let us now examine these plots in more detail.

The prevalence of  $\alpha$ -helices in structures is reflected in the peak at  $\lambda = 3.6$  in Fig. 3(a). As a check, we repeated the analysis for 493 proteins in our database that are classified as mainly  $\beta$  by CATH [20]. The  $\alpha$ -helix peak disappears completely for this subset, and a weaker peak corresponding to  $\beta$  strands at  $\lambda = 2.2$  (which was not visible in Fig. 3(a)) emerges in its place. This may indicate that the formation and arrangement of  $\beta$  strands is less influenced by hydrophobic forces. The other prominent feature of Fig. 3(a) is the increase in  $A_q$  as  $q \rightarrow 0$ . We believe this reflects the fact that at a coarse level the protein is a *compact polymer*; it is well known that polymer statistics leads to long-range correlations in the statistics of segments in the interior of a compact structure [11]. While the precise manner in which this could lead to correlations as in Fig. 3(a) has not been worked out, we note that similar effects have been observed before in studies of protein-like structures in three dimensions [13], and compact lattice polymers in two dimensions [12].

The  $\alpha$ -helix peak, which is prominent in the hydrophobicity power spectrum of Fig. 1(b) is absent from Fig. 3(c). Thus, the observed periodicity in sequence data is not an intrinsic feature of the amino-acid profiles, but dictated by the required folding of structures. If the sequence of amino-acids were totally random, we

would expect a distribution  $P(\{h_i\}) = \prod_i p_a(h_i)$ , where  $p_a(h_i)$  indicates the frequency of a particular base. The corresponding distribution in Fourier space would also be independent of  $q$ . The observed  $\{C_q\}$  are indeed constant (approximately  $0.42 \pm 0.02$ ), at large  $q$ . This constant is different from the average indicated in Fig. 1(b), with the assumption that the amino-acids are distributed randomly. This difference is due to the interaction term in equation 2.

Reduced values of  $C_q$  are observed as  $q \rightarrow 0$ , corresponding to large periodicities, as seen in Fig. 3(c). A similar feature is also present in the power spectrum in Fig. 1(a), as noted before by Irback *et al.* [24] who suggest that anti-correlations can be advantageous for removing the degeneracies of ground state for folding sequences. More recent studies also indicate that long stretches of hydrophobic monomers, which could be a source of long range positive correlations, are avoided [25]. Further investigations of this issue would be helpful.

Finally, we note that the interaction terms  $\{B_q\}$  in Fig. 3(b) which correlate sequence and structure profiles (at different periodicities) are approximately constant. As  $\sum_q \tilde{h}_q \tilde{s}_q^* = \sum_i h_i s_i$ , these terms can be regarded as arising from the Boltzmann weight  $\exp[-E/(k_B T)]$  of a solvation energy  $E = \sum_i h_i s_i$  at some temperature  $T$ . Using  $\overline{B_q} \approx 0.32 \pm 0.03$  kcal/mol, we can extract a corresponding temperature of  $T = (2\overline{B_q})/k_B = 323 \pm 30$ K. Interestingly, this fictitious  $T$  is around room temperature, i.e. in the range of temperatures that most proteins fold and function. This indicates that an important factor in correlating sequence hydrophobicity, and structural solvent accessibility is indeed the free energy of solvation. This conclusion is also consistent with the analysis done by Miller [17, 18], which estimated differences in the free energies of amino-acids between the surface and the core of the proteins by counting their relative frequencies in the different locations.

In principle, the Gaussian distribution in Eq. 2 can be used as a tool for predicting structures, at least as far as their surface exposure profile is concerned. Given a specific sequence, we can calculate the hydrophobicity profile  $\{h_i\}$ , and the corresponding  $\{\tilde{h}_q\}$ . The conditional probability for surface exposure profiles is then given by

$$P(\{\tilde{s}_q | \tilde{h}_q\}) = \prod_q p(\tilde{s}_q | \tilde{h}_q) \propto \\ \prod_q \exp \left[ -\frac{(\Re s_q - \chi_q \Re h_q)^2}{2\sigma_q^2} - \frac{(\Im s_q - \chi'_q \Im h_q)^2}{2\sigma'^2_q} \right]. \quad (3)$$

Thus  $\tilde{s}_q$  is Gaussian distributed with a mean value of  $\chi_q \tilde{h}_q$ , and a variance  $\sigma_q^2$ , with the ‘susceptibility’  $\chi_q$ , and the ‘noise’  $\sigma_q$  easily related to  $(A_q, B_q, C_q)$ . The corresponding distribution of  $\{s_i\}$  in real space is then obtained by Fourier transformation.

We investigated correlations between protein sequences and structures due to hydrophobic forces, by application of Fourier transforms to profiles of hydropho-

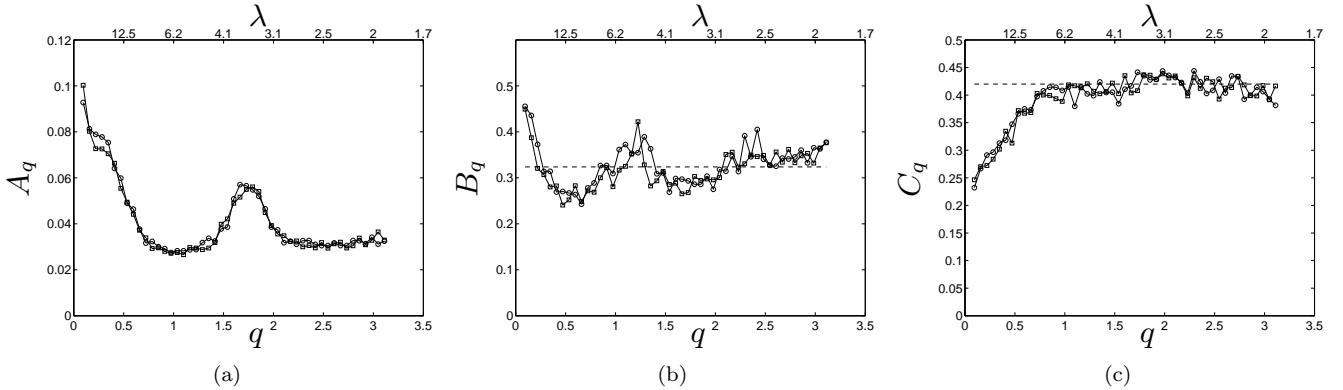


FIG. 3: *Intrinsic* variances of solvent accessibility and hydrophobicity profiles are described by  $A_q$  and  $C_q$  respectively, while  $B_q$  is related to the interaction that correlates them. The square and circle symbols correspond to the parameters of the imaginary and real components, respectively. These figures are calculated for our set of 1461 proteins. Dashed lines indicate respectively the average value of  $B_q$  [in (b)], and the asymptotic behavior of  $C_q$  [in (c)].

bicity and solvent accessibility. Each Fourier component is separately well approximated by a Gaussian distribution; their joint distribution is described by a product of multivariate Gaussians at different periodicities. This approach enables us to separate the intrinsic tendencies of the profiles from the interactions that couple them. We thus find that  $\alpha$ -helix periodicity is a feature of structures and not sequences, and that at long periods the structural profiles are more correlated than average, while the sequences are less correlated. A quite satisfying outcome is that the correlations between the two profiles can be explained by the Boltzmann weight of the solvation energy at room temperatures.

Our joint distribution can be used in applications such as predicting solvent accessibility from hydrophobicity profiles [26], or protein interaction sites [27]. Incorporating the impact of correlations within solvent accessibilities is likely to improve predictions. The distribution can also be used in analytical approaches to protein folding, wherever there is a need for taking into account the complexities of structure and sequence space.

MY and CBB are supported by Functional Genomics Innovation Award (CBB and Philip A. Sharp). MK is supported by NSF through grant No. DMR-01-18213. MY thanks M. Povinelli for helpful discussions.

- [1] W. Kauzmann, *Advan. Protein Chem.* **16**, 1 (1959).
- [2] B. K. Lee and F. M. Richards, *J. Mol. Biol.* **55**, 379 (1971).
- [3] D. Eisenberg, R. M. Weiss, and T. C. Terwilliger, *Proc. Nat. Acad. Sci. US* **81**, 140 (1984).
- [4] S. Moelbert, E. Emberly, and C. Tang, *Protein Sci.* **13**, 752 (2004).
- [5] A. Irback, C. Peterson, and F. Potthast, *Proc. Nat. Acad. Sci. USA* **93**, 9533 (1996).
- [6] V. S. Pande, A. Y. Grosberg, and T. Tanaka, *Proc. Nat. Acad. Sci. USA* **91**, 12972 (1994).
- [7] B. J. Strait and T. G. Dewey, *PRE* **52**, 6588 (1995).
- [8] O. Weiss and H. Herz, *Journal of Theoretical Biology* **190**, 341 (1998).
- [9] F. Eisenhaber, F. Imperiale, P. Argos, and C. Frommel, *Protein-Struct. Funct. Genet.* **25**, 157 (1996).
- [10] J. Wilder and E. I. Shakhnovich, *Phys. Rev. E* **62**, 7100 (2000).
- [11] E. N. Govorun, V. A. Ivanov, A. R. Khokhlov, P. G. Khalatur, A. L. Borovinsky, and A. Y. Grosberg, *Phys. Rev. E* **64**, R40903 (2001).
- [12] M. Yahyanejad, M. Kardar, and C. Tang, *J. Chem. Phys.* **118**, 4277 (2003).
- [13] A. R. Khokhlov and P. G. Khalatur, *Phys. Rev. Lett.* **82**, 3456 (1999).
- [14] K. M. Biswas, D. R. Devido, and J. G. Dorsey, *J. Chromatogr. A* **1000**, 637 (2003).
- [15] S. Rackovsky, *Proc. Nat. Acad. Sci. USA* **95**, 8580 (1998).
- [16] A. Irback and E. Sandelin, *Biophysical Journal* **79**, 2252 (2000).
- [17] S. Miller, J. Janin, A. M. Lesk, and C. Chothia, *J. Mol. Biol.* **196**, 641 (1987).
- [18] A. V. Finkelstein, A. Y. Badrettinov, and A. M. Gutin, *Protein-Struct. Funct. Genet.* **23**, 142 (1995).
- [19] L. Holm and C. Sander, *Nucl. Acid Res.* **26**, 316 (1998).
- [20] C. A. Orengo, A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells, and J. M. Thornton, *Structure* **5**, 1093 (1997).
- [21] J. Fauchere and V. Pliska, *Eur. J. Med. Chem.* **18**, 369 (1983).
- [22] S. J. Hubbard, S. F. Campbell, and J. M. Thornton, *J. Mol. Biol.* **220**, 507 (1991).
- [23] T. R. Weikl and K. A. Dill, *J. Mol. Biol.* **332**, 953 (2003).
- [24] A. Irback, C. Peterson, and F. Potthast, *Phys. Rev. E* **55**, 860 (1997).
- [25] R. Schwartz, S. Istrail, and J. King, *Protein Sci.* **10**, 1023 (2001).
- [26] H. Naderi-manesh, M. Sadeghi, S. Arab, and A. A. Movahedi, *Protein-Struct. Funct. Genet.* **42**, 452 (2001).
- [27] X. Gallet, B. Charlotteaux, A. Thomas, and R. Brasseur, *J. Mol. Biol.* **302**, 917 (2000).